DOCUMENT RESUME

ED 187 761 TM 800 268

AUTHOR Berk, Ronald A.
TITLE A Consumers' Guide to Criterion-Referenced Test "Reliability".
PUB DATE Apr 80
NOTE 61p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Boston, MA, April 8-10, 1980). Tables may be marginally legible due to small print.

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Criterion Referenced Tests; Cutting Scores; Scoring Formulas; *Statistical Analysis; Test Interpretation; *Test Reliability
IDENTIFIERS *Domain Score Estimation; *Loss Function

ABSTRACT

Seventeen statistics for measuring the reliability of criterion-referenced tests were critically reviewed. The review was organized into two sections: (1) a discussion of preliminary considerations to provide a foundation for choosing the appropriate category of "reliability" (threshold loss function, squared-error loss-function, or domain score estimation) for a particular application and (2) an evaluation of the indices for selection of a specific index within the chosen category. The former was related to information on the test forms assumption, the cutting score, probabilities of decision errors, score interpretation, type of decision, and the seriousness of losses associated with the decision errors. The latter suggested that the threshold loss function approaches of Hambleton and Novick (1973) and Huynh (1976) were the most useful, the squared-error loss function indices of Livingston (1972a) and Brennan (1980) provide meaningful information especially for placement tests, and the domain score estimation statistics have limited utility for individual decision making. Detailed tables are included. (Author/GSK)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED187761

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
R. Berk

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

A CONSUMERS' GUIDE TO CRITERION-REFERENCED TEST "RELIABILITY"

Ronald A. Berk[1]

The Johns Hopkins University

Paper presented at the annual meeting of the National Council on Measurement in Education

Boston, April 1980

TM 800268

A Consumers' Guide to Criterion-Referenced Test "Reliability"

Abstract

Seventeen statistics for measuring the reliability of criterion-referenced tests were critically reviewed. The review was organized into two sections: (1) a discussion of preliminary considerations to provide a foundation for choosing the appropriate category of "reliability" (threshold loss function, squared-error loss function, or domain score estimation) for a particular application and (2) an evaluation of the indices to guide the selection of a specific index within the chosen category. The former was related to information on the test forms assumption, the cutting score, probabilities of decision errors, score interpretation, type of decision, and the seriousness of losses associated with the decision errors. The latter suggested that the threshold loss function approaches of Hambleton and Novick (1973) and Huynh (1976) were the most useful, the squared-error loss function indices of Livingston (1972a) and Brennan (1980) provide meaningful information especially for placement tests, and the domain score estimation statistics have limited utility for individual decision making.

A Consumers' Guide to Criterion-Referenced Test "Reliability"

Over the past decade more than a dozen different statistics have been devised for measuring the reliability of criterion-referenced tests. Recent summaries and critiques of those measures have been conducted by Hambleton, Swaminathan, Algina, and Coulson (1978), Linn (1979), and Shepard (1980). Hambleton et al. (1978, pp. 15-23) defined three major categories of reliability: (1) reliability of mastery classification decisions--consistency of mastery-nonmastery classification decision making across repeated measures with one test form or parallel test forms; (2) reliability of criterion-referenced test scores--consistency of squared deviations of individual scores from the cutting score across parallel or randomly parallel test forms; (3) reliability of domain score estimates--consistency of individual scores across parallel or randomly parallel test forms. Subsequently, in-depth analyses of the indices that fall into category one that are based on a threshold loss function and the indices that fall into category two that are based on a squared-error loss function have been presented by Subkoviak (1980) and Brennan (1980), respectively. These sets of indices are appropriate for criterion-referenced tests where mastery-nonmastery decisions are made on the basis of a cutting score. The indices that fall into category three that are used to estimate an individual's domain score or proportion correct have received the least attention.

Despite the number of indices and the available sources noted above, a reliability index is rarely reported for criterion-referenced and minimum competency tests developed by school districts, state departments

of education, and commercial test publishers. When an index is reported, it is usually the Kuder-Richardson 20 or 21. In order to facilitate increased and proper use of the reliability indices by teachers, district and state level test makers, and test publishers, the indices have been compiled and evaluated in the form of a "consumers' guide".

The purpose of this paper is to critically review all statistics that have been proposed as indices of criterion-referenced test reliability. The review will consist of a comparison of the characteristics of the indices and an evaluation of their precision and practicability. This should indicate clearly to a practitioner (1) which category of reliability is most appropriate for a given application and (2) the specific index within the category that should be computed.

Method

A total of 13 different approaches that yield 17 reliability indices for criterion-referenced tests were identified. They were grouped into three categories: (1) threshold loss function, (2) squared-error loss function, and (3) domain score estimation. The selection of a reliability index for a specific application was conceived as a two-stage decision process. Stage one involved the choice of the appropriate category of reliability; stage two involved the choice of a specific index within the chosen category.

The initial choice of a reliability category is based on certain preliminary considerations related to the assumptions, interpretations, and uses of the indices. These considerations are examined first. They include a clarification of terminology, test forms assumption, setting the cutting score, and test score interpretation and decision making.

The second choice of a specific index is based on the characteristics, precision, and practicability of each index. A critical review of the 17 indices according to those factors is presented next. Assembled into "consumers' guide" tables, the indices are scrutinized in terms of their salient characteristics including psychometric definition, number of test administrations, test forms assumption, distribution assumptions, and correction for chance agreement, and evaluated in terms of their advantages and disadvantages with regard to their precision and practicability. Practicability pertained to the ease with which an index could be computed and interpreted.

Preliminary Considerations

A Clarification of Terminology

The use of the term reliability coefficient to characterize the indices recommended for criterion-referenced tests is inappropriate. That term is, in fact, a misnomer since the threshold loss and squared-error loss function approaches are neither defined nor interpreted in the language of the standard psychometric reliability components of observed-score and true-score variance (cf. APA/AERA/NCME, 1974). It is certainly possible to relate many of the indices to classical reliability theory or to describe them within a broad framework of variance components. However, their properties suggest that the term agreement index provides a more accurate label. The threshold loss function indices are measures of agreement between categorical data sets based on mastery-nonmastery classifications; the squared-error loss function indices are measures of agreement between the scores on two test forms. The various indices may be defined and interpreted as measures of overall agreement or as measures of agreement

with the contribution of chance agreement removed.

Unfortunately, the domain score estimation procedures can not be explained as easily in terms of an agreement function. Four of the procedures involve a consideration of different types of standard error, while the others consist of point estimates of the domain score parameter and a reliability-like index. Therefore, the statistics in this last category will not be relabeled.

The preceding change in terminology should provide a more meaningful descriptor that is congruent with the indices' characteristics and will hopefully reduce some of the confusion between those measures and classical reliability coefficients. Hereafter the threshold loss and squared-error loss function indices will be referred to as agreement indices. For simplification, the three categories of threshold loss, squared-error loss, and domain score estimation will fall under the rubric "reliability."

Test Forms Assumption

All of the statistics reviewed in this paper except two point estimates of the domain score require an assumption about test forms. It is assumed that two or more forms of a criterion-referenced test are either classically parallel or randomly parallel. The assumption is related to (1) how the item sample that defines one test form and the samples that define alternate forms are generated and (2) the characteristics of the item domain from which those samples are generated.

Classically parallel test forms. When items are written from objectives-based specifications using traditional item construction rules, those items viewed collectively constitute a sample from a theoretical domain of "all possible items" that could be written for the objective.

When two or more samples are developed for the same objective and domain, those samples should contain content and yield means, variances, and item intercorrelations identical to the first sample. Item samples or alternate test forms with those properties are said to be classically parallel or equivalent to the first form. This statistical equivalence imposes restrictions on the domain. It requires that the content or set of behaviors being measured by the items be relatively homogeneous and, consequently, that the items in the domain possess similar difficulty levels. Very often when a test maker is not able to produce parallel forms, the alternate form(s) and its statistical properties are simply "assumed." All of the agreement indices except Brennan's (1980) index and one domain score estimation statistic (Lord & Novick, 1968) are based on this assumption.

Randomly parallel test forms. In contrast, the recent technological advances in the specification of behavioral domains offer alternative strategies that can generate "all possible items" in a domain. The progress that has been made over the past decade is reflected in the methodologies of amplified objectives, IOX test specifications, mapping sentences (facet theory), item transformations, item forms, and algorithms (cf. Millman, 1980; Popham, 1980). Proponents of the latter three strategies, in particular, claim that generating an item domain is an objective, mechanical process (Berk, 1979, in press a). While the precision of the different strategies varies markedly and much more research lies ahead, the strategies do provide some evidence that the skepticism expressed by Thorndike in 1967 was not totally warranted, "As soon as we try to conceptualize a test score as a sample from some universe, we are brought face to face with the very knotty problem of defining the universe from which we are sampling (p.285)...the universe is considerably restricted, is hard to define,

and the sampling from it is hardly to be considered random" (p.288).

When one of the strategies is employed in test construction, the items to be included on one test form or alternate test forms are selected from the item domain using a random or stratified random sampling plan. These test forms are said to be randomly parallel. None of the statistical properties mentioned previously apply to these forms nor is there any condition imposed that limits the variability of item difficulty levels in the domain. When the test maker can not actually build test forms by randomly sampling from the domain, the alternate form(s) and its characteristics can be "assumed" (cf. Cardinet, Tourneur, & Allal, 1976). As Brennan (1980) indicates:

> . . . the assumption of randomly parallel tests is a useful idealization. Although the items of a particular test may not have been drawn at random, one can usually conceive of a universe of items from which the test items might have been drawn. (p.226)

Only one agreement index (Brennan & Kane, 1977a) and four domain score estimation statistics (Berk, in press b; Brennan, 1980; Lord, 1957) assume randomly parallel test forms.

Setting the Cutting Score

Computation of the threshold loss and squared-error loss function indices presumes a cutting score for mastery of an objective has been selected. All of the indices are sensitive to the position of the cutting score in the score distribution. This relationship makes it

essential that the interpretation of an index includes a specification of the cutting score. Therefore, the setting of standards is a crucial step prior to the consideration of an agreement index.

Setting standards is one of the most important, controversial, and misunderstood validity issues in criterion-referenced measurement. A decision maker must confront the accuracy of mastery-nonmastery classification decisions. Evidence of accuracy expressed usually as probabilities of correct and incorrect (false mastery and false nonmastery) classifications pertains to decision validity (Hambleton, 1980, pp. 97-99). Since an extensive discussion of this topic is not within the scope of this paper, interested readers are referred to Hambleton (1980) and Hambleton, Powell, and Eignor (1979) for reviews of methods for setting cutting scores and estimating probabilities.

The major point of this section is that without validity evidence or a sound justification for setting the cutting score it seems pointless even to compute an agreement index. Certainly one can compute an index based on a cutting score that has little or no justification. Since the final interpretation links the index value to the cutting score, a high agreement index associated with an "invalid" or "unjustified" standard, for example, might indicate that a test can consistently classify students into the wrong groups. Consistent decision making without accurate decision making has questionable value in criterion-referenced evaluation.

In addition, a priori decisions regarding the relative seriousness of the losses related to the probabilities of false mastery and false nonmastery errors must be reached before choosing a loss function. Whether such losses are perceived to be equally serious (threshold loss function) or to vary as a function of how far above or below from the

cutting score the misclassified students' scores are located (squared-error loss function) will be the determining factors in that choice.

Test Score Interpretation and Decision Making

The scores on a criterion-referenced test can be used for decision making about individuals and programs at the classroom, school, district, state, and national levels. The decisions made by the classroom teacher within the context of mastery learning theory and diagnostic-prescriptive teaching have best demonstrated the utility of criterion-referenced test scores. Typically, those decisions are derived from an objectives-based assessment of individual strengths (mastery) and weaknesses (non-mastery). They involve placing students at the appropriate points in the instructional sequence, monitoring students' successes and failures as they progress through the program, and assigning letter grades/pass-fail to certify competencies in a subject area or course. Recently the latter type of summative evaluation has been extended in the form of minimum competency certification for grade-to-grade promotion and high school graduation. Even outside the public school arena, similar score interpretations are being contemplated for the certification and/or licensing of professionals in medicine and the allied health fields, social work, psychology, law, and engineering.

In addition to these different types of individual decisions, the scores may be used to assess program effectiveness at the decision levels mentioned above. These evaluation decisions focus usually on the effectiveness of curricular materials, instructional methods, instructional media, or any combination of materials, methods, and media. The

actual decisions may be to continue, to modify, or to terminate a given program. In order to supply information that will be meaningful for these decisions it is often necessary to summarize the test score results into various forms such as: (1) percentage of students demonstrating mastery of each objective, (2) percentage of students demonstrating mastery of sets of objectives, (3) estimated mean performance on the total test, and (4) estimated mean performance based on the matrix sampling of students and items from their respective populations or domains. The appropriateness of these summaries will vary as a function of the level of the decision maker, from classroom teacher to state superintendent, and the nature of the decision. The summary that is needed for a particular decision is composed by aggregating the individual test scores at the appropriate decision level. Examples of several report formats have been assembled by Premer (1978).

The initial choice of a "reliability" category is contingent upon the intended interpretation of the test scores, type of decision to be made with those scores, and the consequences or losses associated with false mastery and false nonmastery decision errors. A brief list of score and decision characteristics related to the three categories of "reliability" is shown in Table 1. Since the score interpretations for the numerous levels of program evaluation decisions are derived from aggregates of individual scores and their corresponding uses and interpretations, only individual score and decision characteristics are specified in the table. A detailed explication of these factors in conjunction with the categories of "reliability" and the indices within each category is presented in the "Review" sections that follow.

Insert Table 1 about here 12

Review of Threshold Loss Agreement Indices

The concept of classification decision consistency in terms of the threshold loss function was first proposed by Hambleton and Novick (1973, p. 168). The use of the threshold loss function assumes (1) a dichotomous, qualitative classification of students as masters and nonmasters of an objective based on a threshold or cutting score, and (2) the losses associated with all false mastery and false nonmastery classification errors, regardless of their size, are equally serious.

The characteristics and evaluation of the seven approaches in this category are outlined in Table 2. With the exception of the Carver (1970) approach that is not sensitive to classification consistency, the various approaches actually involve only two agreement indices: p_o, proportion of individuals consistently classified as masters and nonmasters across (classically) parallel test forms and κ, proportion of individuals consistently classified beyond that expected by chance.

Insert Table 2 about here

---

p_o Versus κ

The choice between p_o and κ as indices of classification consistency depends upon how consistency is defined in a given decision context and how the properties of the statistics are evaluated. This comparison will concentrate on the appropriateness of the indices for tests used for individual decisions at the classroom level, individual certification decisions at the school level (e.g., minimum competency testing), and program evaluation decisions at the district level.

Advantages of p_o. The p_o index measures the overall consistency of mastery-nonmastery classifications. Subkoviak (1980, p. 152) noted two factors that contribute to that consistency: (1) the mastery-nonmastery composition of the group tested and (2) the precision or accuracy of the test itself. The index is sensitive to the selected cutting score, test length, and score variability. The position of the cutting score tends to have a much more profound effect upon the magnitude of the index than either of the other two characteristics. Higher values of p_o correspond to cutting scores at the tails of a unimodal score distribution and lower values correspond to cutting scores near the mean (Subkoviak, 1980). This trend may not follow for bimodal distributions. The index values also increase as the number of test items increases and the score variance increases. In practice, however, it is possible to obtain p_o values of .75 or higher with subtests containing fewer than 10 items and exhibiting relatively little score variance. These properties are especially important in view of the structure and uses of most teacher-made criterion-referenced tests.

Preliminary evidence on the single test administration estimates of p_o (Huynh, 1976; Marshall, 1976; Subkoviak, 1976) also seems to suggest that this index can be estimated with precision for most unimodal distributions (See, for example, Subkoviak, 1980). In one investigation of Huynh's (1976) beta-binomial model it was found that p_o could be estimated with only about 2.5% negative bias (Huynh & Saunders, 1979).

Although these advantages indicate why p_o might be the preferred index of agreement for tests intended for individualized instructional decisions at the classroom level, the disadvantages associated with the uses of κ will demonstrate more dramatically why κ may not be the preferred index for such tests. The rejection of p_o in favor of Cohen's (1960) κ coefficient was advanced originally by Swaminathan, Hambleton, and Algina (1974, p. 264). They felt that κ was somewhat more appropriate since it took into account the agreement that could be expected by chance alone. The κ index measures the test's contribution to the overall proportion of consistent classifications, that is, test consistency. It is calculated by subtracting from p_o the proportion of consistency expected from or attributable to the particular mastery-nonmastery composition of the group (Subkoviak, 1980). The properties of κ that stem from this "correction for chance agreement" are what make this index problematic.

Disadvantages of κ. There are three properties of the κ index that render it undesirable for criterion-referenced tests used for classroom decision making: (1) the correction for chance is constrained by the marginal frequencies and (2) and (3) κ's sensitivity to test length and test score variability may produce restrictions in range. All three properties complicate the interpretability of κ.

First, the correction for chance agreement is constrained by the marginal frequencies of the particular 2 X 2 contingency table data set under consideration. An index of +1.00 can be obtained only when the marginals are equal, i.e., 50% masters and 50% nonmasters on both forms or measurements. Livingston and Wingersky's (1979) arguments on this issue are most persuasive:

> Applying such a correction to a pass/fail contingency table is equivalent to assuming that the proportion of examinees passing the test could not have been anything but what it happened to be. For example, if 87% of the examinees passed the test, kappa will "correct for chance" under the assumption that "chance" would result in exactly 87% of the examinees passing the test. This assumption makes sense when the pass/fail cutoff is chosen on the basis of the scores to which it will be applied, so as to pass a specified proportion of the examinees. It does not make sense when the pass/fail cutoff represents an absolute standard that is to be applied individually to each examinee. (p.250)

One condition under which the correction could be justified is where the selection of the cutting score is determined, in part, by the consequences (losses) of passing or failing a particular proportion or segment of the population. This cutting score is "relative" rather than "absolute." Exemplary of this type of cutting score are the performance standards for minimum competency tests that are set by district or state level decision makers. When these tests are used for individual certification decisions, the cutting score is usually adjusted according to the political, economic, social, and/or instructional consequences of not passing or certifying a certain proportion of the students or, more specifically, of a minority group in the school district.

Second, the values of $\hat{\kappa}$ increase with test length. This is troublesome in criterion-referenced testing because an item sample or subtest designed to assess mastery of an instructional objective consists usually of between three and 10 items on tests developed for classroom level usage and of rarely more than 20 items on tests developed for district or state level usage. For the majority of test applications at these levels, then, the values of $\hat{\kappa}$ may be restricted in range due to short subtest lengths. This restriction would not be of concern in large-scale minimum competency testing programs where the items are often aggregated across objectives to apply a single cutting score to the total test and, subsequently, to estimate classification consistency based on that score.

Third, the values of $\hat{\kappa}$ increase with score variability. A restriction in variance that would be accompanied by a restriction in the range of $\hat{\kappa}$ will occur frequently in criterion-referenced testing due to short subtests as discussed above and to the homogenous composition of the group that is tested. In many cases the test results tend to yield either a large proportion of masters or nonmasters. This is especially evident when classroom test scores are used for individual placement, formative, and summative decisions. The former two types of decision that are typically diagnostic in nature and made prior to instruction can produce test results with disproportionately large numbers of nonmasters; the latter type of decision made at the conclusion of instruction will generate (hopefully) results with disproportionately large numbers of masters.

In addition to these properties of κ, there are other characteristics worthy of consideration: (1) κ's sensitivity to the cutting score is just the inverse of p_o, such that higher values of κ correspond to cutting scores near the mean score and lower values correspond to cutting scores at the tails of the distribution (Huynh & Saunders, 1979; Subkoviak, 1980), (2) $\hat{\kappa}$ is a biased estimator (Harris & Pearlman, 1977, p.35), (3) one of the single administration approaches (Huynh, 1976) provides less accurate estimates of κ (about 10% error) than of p_o, and (4) the interpretation and usefulness of the κ index in practice is not completely clear (See, for example, Safrit & Stamm, 1979).

Recommendations. The foregoing comparison of p_o and κ as indices of classification consistency suggests the following guidelines: (1) p_o should be the index of agreement for criterion-referenced tests used by classroom teachers where an "absolute" cutting score is chosen and for other tests that may contain short subtests and/or yield low score variance; (2) κ should be the index of agreement for criterion-referenced tests used for individual certification decisions and program evaluation decisions where "relative" cutting scores are set according to the consequences of passing or failing a particular proportion of the students and a restriction in range due to test length or score variability is not anticipated. Given the dependence of both indices on the cutting score, test length, score variability, and the cell frequencies of the 2 X 2 contingency table, it is recommended that the test maker report all of this information along with the agreement index to facilitate perceptive and correct index interpretations (cf. Livingston & Wingersky, 1979; Millman, 1974).

One- Versus Two-Administration Approaches

Among the six classification consistency approaches to estimating p_o and κ, the Hambleton and Novick (1973) method appears to have the greatest utility for classroom test construction and the Huynh (1976) procedure has definite potential for use with tests developed by evaluators at the district and state levels and by test publishers.

Hambleton and Novick method. The Hambleton and Novick method provides the only unbiased estimate of p_o. It is also the easiest method to understand, to compute, and to interpret. Unfortunately, there are two problems that must be confronted: (1) $\hat{p}_o$ has comparatively large standard errors (6 to 8%) for classroom-size samples (n=30) (Subkoviak, 1980) and (2) given that two test administrations are required, the teacher must either develop parallel forms or deal with possible testing effect bias as a result of the test-retest design.

With regard to the accuracy of the $\hat{p}_o$ estimate, teachers should be urged to interpret the index values with caution. Since the preliminary evidence indicates that the error decreases as the cutting score increases on short subtests, a consideration of about 6% error in evaluating classification consistency based on mastery standards at or above 80% seems advisable.

The requirement of two repeated measurements with one test form or two parallel forms to estimate p_o has been judged as a disadvantage of the Hambleton and Novick method. The problem with the two administrations at the classroom level is not the inconvenience of the administrations per se, as a considerable amount of pre- and posttesting are common practice as integral components of the teaching-learning process, but the construction of parallel test forms. It is quite difficult and very time consuming to build two classically parallel

tests for a given set of objectives. To circumvent this problem it is recommended that the $\hat{p}_o$ estimate be computed from repeated measures of one test. For many content domains the traditional weakness of testing effect associated with the test-retest design will be negligible. When a significant testing effect is suspected, the teacher should be cognizant of its influence on the decision-consistency estimates. Since knowledge gained from the initial testing tends to spuriously inflate the scores on the second testing, only the status of nonmasters on testing one will be altered on testing two. If several nonmasters become masters due to testing effect, it is likely that $\hat{p}_o$ will be spuriously deflated. This conservative bias should also be taken into account in interpreting the $\hat{p}_o$ index for teacher-made criterion-referenced tests.

Huynh method. In lieu of the above method, the more sophisticated test maker at the district or state level should seriously consider the Huynh procedure. Subkoviak (1980) has recommended this approach on statistical as well as practical grounds. It provides relatively precise though conservatively biased estimates of p_o and κ, even for classroom-size samples. While the beta-binomial model is one of the most conceptually, mathematically, and computationally complex threshold loss function approaches, it has several distinct practical advantages: (1) it requires only one test form and administration, (2) a few recent studies indicate that it is sufficiently robust when the assumption of equal item difficulties is violated in practice (Huynh & Saunders, 1979; Subkoviak, 1978), (3) tabled values of $\hat{p}_o$ and $\hat{\kappa}$ and their standard errors have been prepared for subtests of between five and 10 items related to an objective based on four pieces of information -- number of items, cutting score, mean score, and KR-21 reliability co-

efficient for the subtest (Huynh, 1979; Subkoviak, 1980), and (4) a computer program with documentation is available to calculate the indices for all other size subtests (Huynh, 1979). Furthermore, Huynh's model which is intended for dichotomous items has been extended to the case of multipoint items in order to apply the methodology to criterion-referenced measures of psychomotor and affective behaviors (See Berk, 1980).

Review of Squared-Error Loss Agreement Indices

In contrast to the preceding loss function which focuses on the "consistency of classifications," the squared-error loss approach to measurement error deals with the "consistency of measurements or scores." It is based on the squared deviations of individual scores from the cutting score. This builds in a sensitivity to degrees of mastery and nonmastery along the score continuum, not just the qualitative master-nonmaster classification assumed by the threshold loss function. That sensitivity means that a squared-error loss agreement index would reflect the magnitude of all errors of measurement including those that do not lead to misclassification (Brennan & Kane, 1977a, p. 287). Furthermore, the losses associated with false mastery and false nonmastery errors are not assumed to be equally serious. The consequences of misclassifying students who are far above or below the cutting score are considered more serious than the consequences of misclassifying those who are close to the cutting score.

$k^2(X,T_x)$ Versus $\Phi(\lambda)$

An evaluation of the two agreement indices in this category is shown in Table 2. The indices are quite similar in regard to what they are measuring, although the actual formulations are very different. The fundamental distinctions between the indices pertain to their assumptions about test forms (classically parallel or randomly parallel) and, based on these assumptions, their definitions of error variance or squared-error loss.

Insert Table 3 about here

Brennan (1980) defined two types of error variance in terms of analysis of variance components: $\sigma^2(\delta)$, person X item interaction and $\sigma^2(\Delta)$, item effect plus person X item interaction. The former, which is derived from the assumption of classically parallel test forms, is part of Livingston's (1972a) $k^2(X,T_x)$ index; the latter, which is derived from the assumption of randomly parallel test forms, is incorporated in Brennan's (1980) $\Phi(\lambda)$ index.

There are several technical features common to both indices that should be mentioned: (1) no distribution assumptions are necessary, (2) as agreement indices, they are uncorrected for chance agreement (See Kane & Brennan, in press), (3) index values can change, independent of their standard errors of measurement, (4) index values increase as the cutting score is set farther and farther from the mean (similar to values of p_o), (5) index values increase as test length (number of items sampled) increases due to decreases in error variance, and (6)

when there is no true or universe score variance, the indices can still have positive values as long as the true score mean does not equal the cutting score. It is also interesting to note that when the cutting score equals the sample mean, $\hat{k}^2(X,T_x)$ reduces to the KR-20 coefficient and $\hat{\Phi}(\lambda=\bar{X})$ is identical to the KR-21 coefficient.

Criticisms of $k^2(X,T_x)$. Criticisms of some of these technical features of Livingston's (1972a) index have been expressed by Harris (1972), Shavelson, Block, and Ravitch (1972), and Hambleton and Novick (1973). Since the criticisms may also be directed at Brennan's (1980) index, the responses that follow apply to both k^2 and $\Phi(\lambda)$.

Harris (1972) argued that "although Livingston's reliability coefficient is (generally) larger than the conventional one, the standard error of measurement is the same, and consequently this larger coefficient does not imply a more dependable determination of whether or not a true score falls below (or exceeds) a given criterion value" (p.29). Brennan and Kane (1977a, p.286) felt that this criticism was primarily a matter of "form rather than substance." Livingston's k^2 index and the standard error of measurement are different statistics that provide different information for different purposes. The k^2 index as well as the $\Phi(\lambda)$ index measure the consistency of scores in relation to the cutting score on repeated testings; the standard errors of measurement $\sigma(\delta)$, which is identical to that derived from classical reliability theory, and $\sigma(\Delta)$, respectively, measure the inconsistency or average imprecision of an individual's scores independent of the cutting score on repeating testings. These standard errors could be used as domain score estimation statistics for

setting up confidence intervals around individual scores. Those statistics do not diminish the meaningfulness and utility of k^2 and $\Phi(\lambda)$. As Livingston (1972b) pointed out in his reply to Harris, "...reliability is not a characteristic of a single score, but of a group of scores...the larger criterion-referenced reliability coefficient does imply a more dependable overall determination of whether each true score falls above or below the criterion level, when this decision is to be made for every individual score in the distribution" (p.31).

Shavelson et al.(1972) cited three problems with k^2: (1) it is "a function of the criterion as well as a function of individuals' responses to items...[consequently, it] is not directly related to the repeatability of the measure" (p.134), (2) "this theory is applicable to situations where there may be no differences among true scores of individuals...this seems unnecessary" (pp.134-135), and (3) $\hat{k}^2$ "does not estimate the ratio of true score variance over observed score variance alone; $k^2(X,T_x)$ cannot be used or interpreted like conventional reliabilities...we suggest that this statistic be given some other name than 'reliability' "(p.135).

The first charge pertains to the sensitivity of the indices to the cutting score, specified previously as technical feature (4). Livingston (1972c) argued that "It is entirely appropriate that [k^2]... should depend heavily on the difference between the group mean score and the criterion score...[this] difference...can be considered 'true variance' for criterion-referenced purposes" (p.139); Brennan (1979) contended that "As an examinee's observed score gets farther away from

the cutting score, it becomes less likely that the examinee will be misclassified... Since 'most' examinee scores are concentrated around the mean, it seems intuitively reasonable that $\Phi(\lambda)$ should increase as the cutting score moves away from the mean" (p.38). Both researchers perceive the characteristic as desirable rather than as a limitation. However, it is essential that the test maker report the cutting score upon which a particular index value is based along with other information so that the relationship is considered in interpreting the magnitude of $\hat{k}^2$ and $\hat{\Phi}(\lambda)$. Brennan (1980, p.228) also suggests that it might even be better to report a "curve" of index values as a function of several possible cutting scores. Furthermore, the Shavelson et al. (1972) assertion regarding the concept of repeatability does not appear to follow from the dependence of the index on the cutting score. Both k^2 and $\Phi(\lambda)$ are defined in terms of the repeatability of squared deviations of scores from the cutting score rather than from the mean score.

The second charge focuses on the circumstances whereby the indices may yield positive values when there may be no true score variance, specified previously as technical feature (6). This property does not render the index as "unnecessary." Livingston (1972a, p.19) noted that even if a group of students randomly guessed at all items and the resultant set of scores produced a high index, that index would be meaningful; the test user could be quite confident that all of the students had true scores below the cutting score. The

index conveys what proportion of the information provided by the test is reliable information (Livingston, 1972c, p.139). Brennan (1980) offered an explanation for the situation where the error variance is positive and all students have identical high universe scores compared to the cutting score. He indicated that "In this case, $\Phi(\lambda)$ will be positive, reflecting the fact that it is relatively easy to determine correctly whether or not examinees are above the cut-off score" (p.203).

The third charge relates to the appropriateness of the term "reliability" for describing and interpreting the indices. Livingston's (1972c) position on this issue is clear, "[k^2] deserves to be called a reliability coefficient because it represents the ratio of 'true' to 'observed' mean squared deviations from the criterion score, and the criterion-referenced correlation between alternate forms of the same test, and the squared criterion-referenced correlation between true scores and observed scores" (p.139). Brennan and Kane (1977a, p.285) have rejected the term "reliability coefficient" in favor of another term "dependability coefficient" for three reasons: (1) $\Phi(\lambda)$ is defined in terms of expected squared deviations rather than variances, (2) $\Phi(\lambda)$ entails definitions of error and parallel tests not usually associated with reliability coefficients, and (3) the magnitude of $\Phi(\lambda)$ depends upon a constant (λ), while the magnitude of the error variance is independent of λ. My own preference for the term "agreement index" and the rationale for that preference which is similar, in part, to Brennan and Kane's first reason were explicated at the beginning of this paper.

Unlike the foregoing criticisms by Harris (1972) and Shavelson et al. (1972) that concentrate on technical characteristics of the indices, Hambleton and Novick's (1973) comments address the more fundamental issue of whether the squared-error loss approach to estimating "reliability" should be employed at all in criterion-referenced measurement. They stated emphatically that "Livingston misses the point for much of criterion-referenced testing...the problem is one of deciding whether a student's true performance level is above or below some cutting score...Livingston's choice of a loss function...is wrong" (p.168). Recently Hambleton et al. (1978, p.17) admitted that their position is based on "conjecture" and only time will tell if they are correct. Given the numerous decision applications and interpretations of criterion-referenced test scores that have been discussed in preceding sections, Hambleton and Novick's (1973) recommendation that only the threshold loss approach is appropriate would appear to be overly restrictive and, generally, imperceptive.

The guidelines presented previously for determining the appropriate loss function dealt with the type of decision to be made with the scores and the losses associated with the two types of decision error. There are some contexts where one of the loss functions may be preferable and other contexts where neither may be preferable. Shepard (1980) suggested an "ideal index" that implies another loss function; it would reflect only false mastery and false nonmastery classification errors and it would weight those errors by degree. Unfortunately, at present such an index and loss function do not exist.

All of the aforementioned criticisms of Livingston's (1972a) index, viewed retrospectively, seem to have emanated out of the con-

fusion and misinterpretations related to criterion-referenced measurement in the early 1970s and from the mistaken efforts to develop a "reliability theory" for a new criterion-referenced testing methodology within the delimited framework of classical reliability theory. The methodological changes over the past decade preceeded rapidly and the insights into what "reliability" means in criterion-referenced measurement at this date will hopefully be more perceptive.

Due to the recency of Brennan's (1977a) formulation and, perhaps, to a greater understanding of the methodology, no formal criticisms of the $\Phi(\lambda)$ index have appeared in the literature. Instead, the few comments that have focused on his approach (See, for example, Linn, 1979; Shepard, 1980) have been quite laudatory.

Advantages of $k^2(X,T_x)$ and $\Phi(\lambda)$. On the positive side, both indices have certain advantages that deserve serious attention by test makers at the classroom, district, and state levels. First, the estimates $\hat{k}^2$ and $\hat{\Phi}(\lambda)$ require only one test administration. Second, the indices have particular utility for criterion-referenced tests used by teachers to place students into instructional treatments. Once an initial assessment of mastery-nonmastery status has been completed based on the cutting score, the information typically of greatest value to teachers prior to instruction is the degree of mastery or nonmastery of each student along the score continuum. The detailed diagnosis of how much and what content a student does or does not know can assist the teacher in planning and assigning individualized prescriptions. The degree of mastery would suggest the nature of enrichment activities or the appropriate instructional strategies and materials for the next objective; the degree

of nonmastery which may indicate the specific aspects of the objective that were not mastered would suggest the nature of remedial activities. This test application and type of decision making is consistent with the technical features of the squared-error loss function. Third, from a practical standpoint, the ease with which both indices could be manually calculated is comparable to that of the KR-20. The formula and computational examples for $\hat{\Phi}(\lambda)$ prepared especially for practitioners by Brennan (1980, pp.211-212) should contribute greatly to that ease. For test makers with computer resources, computer programs with extensive documentation are also available (e.g., Brennan, 1979; Erlich & Shavelson, 1976).

In addition to the above advantages, Brennan's (1980) approach has the advantage of flexibility. The $\Phi(\lambda)$ index is derived from Cronbach, Gleser, Nanda, and Rajaratnam's (1972) generalizability theory. This theory employs analysis of variance designs to estimate "generalizability coefficients" and the magnitude of individual variance components. It permits considerable flexibility in the analysis of different test structures corresponding to different test score uses. Brennan's application of the theory to criterion-referenced measurement makes it possible to estimate $\Phi(\lambda)$ and universe score variance and error variance components so that they reflect the various facets of interest in a given situation. He has illustrated this flexibility with a test composed of items randomly sampled from a stratified as opposed to an undifferentiated domain (See Brennan, 1980, pp.221-223).

Recommendations. With regard to the interpretation of the indices, it is strongly recommended that the cutting score, test length, and estimated error variance [either $\hat{\sigma}^2(\delta)$ or $\hat{\sigma}^2(\Delta)$] be reported to test users. This information should furnish an adequate base for index interpretation given the number and diversity of factors that affect the values of $\hat{k}^2$ and $\hat{\Phi}(\lambda)$. Brennan (1980) also urges test makers to report estimates of universe score variance and the squared deviation of the grand mean from the cutting score and, whenever possible, to estimate $\Phi(\lambda)$ in terms of variance components. This emphasis on variance components is consistent with the guidelines set forth in Standards for Educational & Psychological Tests (APA/AERA/NCME, 1974):

> The estimation of clearly labeled components of score variance is the most informative outcome of a reliability study, both for the test developer wishing to improve the reliability of his [or her] instrument and for the user desiring to interpret test scores with maximum understanding. (p.49)

Finally, the choice between $k^2(X,T_x)$ and $\Phi(\lambda)$ is reduced to the simple decision of whether the assumption of classically parallel test forms or randomly parallel test forms, respectively, is appropriate for the test that was constructed.

Review of Domain Score Estimation Statistics

Unlike the agreement indices reviewed in the previous sections, the domain score estimation statistics are concerned generally with estimating an individual's score or proportion correct in the item domain

independent of any mastery standard. Seven statistics in this category are described in this section. All of them have meaning for individual decision making at the classroom or school level. Program evaluation decisions would require estimates of the average domain score and the standard error of that average. These estimates vary as a function of the multiple matrix sampling design that is used in the school district. A discussion of the different estimates corresponding to the different matrix sampling designs is beyond the scope of this review. Interested readers should consult Sirotnik's (1974) chapter which is oriented especially toward the practicing evaluator. More technical treatments of the topic include Lord and Novick's (1968, Chap. 11) chapter and Shoemaker's (1973) monograph.

The seven domain score estimation statistics are evaluated in Table 4. They are organized into two subcategories--individual specific and group specific. The characteristics of these subcategories will be examined first.

Insert Table 4 about here

Individual Specific Versus Group Specific

The individual specific statistics are defined, computed, and interpreted separately for each individual. They consist of two point estimates of the proportion of items an individual can answer correctly in the item domain and two estimates of standard error that can be used to set up a confidence interval around each individual's observed proportion correct score. The group specific statistics represent averages of particular individual specific statistics over persons. Two of those statistics consist of estimates of the standard error of measurement based on classi-

cally parallel and randomly parallel test forms' assumptions. The third statistic is an agreement index devised exclusively for domain-referenced interpretations.

Distribution assumptions. Among the seven statistics only the point estimate $\hat{\gamma}_i$ requires distribution assumptions. That approach assumes a beta distribution of true domain scores and a normal prior distribution for which the mean and variance can be specified exactly. The latter assumption may not often be supported in practice.

While the four standard errors ϵ_p, $S.E._{meas.}(\underline{x}_{\underline{a}})$, σ_E, and $\sigma(\Delta)$ can be estimated without any distribution assumptions, some assumptions are necessary when those errors are used to provide interval estimates of the domain score. A distribution assumption is essential for each standard error of measurement in this application so that a probability level can be assigned to the established confidence interval (e.g., 68%, 95%). Therefore, it is recommended that the binomial distribution be assumed for the two individual specific standard errors, the normal distribution be assumed for the traditional $\hat{\sigma}_E$, and either the binomial or compound binomial distribution may be assumed for $\hat{\sigma}(\Delta)$.

Point estimates. The two point estimates of the domain score constitute the simplest and the most complex individual specific approaches. The simplest is $\hat{p}$, the observed proportion of items answered correctly by an individual. This estimate of the domain score P is unbiased and highly reliable for full length tests. However, for short subtests that are keyed to instructional objectives the estimate can be very unreliable. The most complex estimate $\hat{\gamma}_i$ that involves a quasi-Bayesian approach has been demonstrated to yield more precise values than $\hat{p}$ and several alternative

classical and Bayesian estimates even for subtests with as few as eight items (See Hambleton, Hutton, & Swaminathan, 1976, for details). The trade-offs the test maker must consider to obtain this increase in precision over $\hat{P}$ are the conceptual and computational complexity of the model and some distribution assumptions that may be rather tenuous in many practical applications.

Interval estimates. Four of the remaining five approaches consist of different types of standard error that can be used to set up a confidence interval or band within which the domain score P lies. All of the estimates of error measure the degree of imprecision or inconsistency of an individual's scores on repeated testings.

The two individual specific estimates $\hat{\epsilon}_p$ and $S.E._{meas.}(x_a)$ are identical when the item domain is assumed to be infinite. The former is defined in the language of survey sampling theory where an individual's test performance is expressed as a "p value" or the percentage or proportion of items answered correctly (p); the latter is defined in the language of psychometric theory where an individual's test performance is expressed as a raw score (x_a). Both are based on the binomial formula for the "... sampling error of the number of white balls in a sample drawn at random from an urn containing a large number of white and black balls" (Lord, 1957, p. 511). The equivalence of the two unbiased estimates is readily apparent from their computational formulas:

$$\hat{\epsilon}_p = \sqrt{\frac{pq}{n-1}} \qquad (1)$$

and

$$S.E._{meas.}(x_a) = \sqrt{\frac{x_a(n - x_a)}{n-1}} \qquad (2)$$

The formulas recommended by Berk (in press b) and Millman (1974b) include the finite population correction (fpc) that is appropriate for finite item domains. It is shown below:

$$\hat{\varepsilon}_p = \sqrt{\frac{(N - n)}{N}\left(\frac{pq}{n - 1}\right)} \qquad (3)$$

In most situations, however, formula (1) can be employed for infinite as well as finite domains as long as the projected or real domain size (N) is large relative to the sample size (n).

The application of this standard error of measurement in the context of an interval estimate can focus on either an individual's P or the P standard for mastery. Establishing a confidence interval or Uncertainty Band (Millman, 1974b) around the cutting score is one method for relating the standard error to mastery-nonmastery decisions. Individuals with scores that lie above the error region and those with scores that lie below may be classified as masters and nonmasters, respectively. If a 95% confidence interval is used, the probability of correct mastery and nonmastery decisions will be at least 95% (See Millman, 1974b, for some examples). This concept can also be linked to the issue of test length. The standard error could be specified a priori to guide the determination of the number of items to be randomly sampled from a finite item domain (See Berk, in press b).

Probably the most distinctive feature of the individual specific estimate is that for any given value of P, the error remains the same for all tests of the same length irrespective of item content, format, or statistical characteristics (Lord, 1957, 1959; Swineford, 1959). Unfortunately, due to the fact that the estimate is a function of only $\hat{P}$ and the number of items, short subtests, say 10 items or less, measuring

particular objectives can reduce the accuracy of the $\hat{P}$ estimate that, in turn, will diminish the interpretability of the standard error of measurement. Therefore, it is recommended that longer subtests or the total test be used to estimate the error.

The two group specific estimates $\hat{\sigma}_E$ and $\hat{\sigma}(\Delta)$ are averages of the individual specific errors of measurement $\hat{\sigma}(E_{\underline{pa}})$ and $S.E._{meas.}(\underline{x}_{\underline{a}})$, respectively. These relationships have been proven by Lord and Novick (1968) and Brennan and Kane (1977b). Since the single average or "blanket estimate" of the standard error of measurement is applied to each individual score, the ideal level of precision is attained only when each individual has the same error. This seldom occurs in practice. Consequently, $\hat{\sigma}_E$ and $\hat{\sigma}(\Delta)$ should be considered as rough approximations of their corresponding individual specific errors.

These approximations are affected by test length and by the composition of the group tested or form of the score distribution. Based on the previous discussion, if the individual specific estimates of error are markedly influenced by the number of items, so are the averages. Therefore, it is advisable to use subtests with at least 10 items or, even better, to use the total test to obtain reliable estimates of group specific error. With regard to the score distribution, the best approximations seem to require homogeneous group performance. If the distribution contains a large number of extremely high and/or extremely low scores plus some moderate scores as in the case of a skewed or bimodal distribution, the estimate of error would be comparatively low. This result is misleading because that estimate, when applied to each of the scores, would tend to overestimate the standard errors for the extreme scores and underestimate the errors for the moderate scores. One way to minimize and possibly eliminate this biase is to employ a homogeneous data base. That can be

accomplished, using Lord and Novick's (1968, p. 155) suggestion, by partitioning a given score distribution (whether it is normal, bimodal, or another shape) into sections that identify subgroups of relatively similar ability and then computing $\hat{\sigma}_E$ or $\hat{\sigma}(\Delta)$ for each subgroup. This strategy would assure that each estimated subgroup standard error of measurement will be generally uniform for all points in that part of the distribution to which it is applied.

Selection of a standard error of measurement. The choice among the individual specific standard error of measurement $S.E._{meas.}(\underline{x}_a)$ and the two group specific standard errors $\hat{\sigma}_E$ and $\hat{\sigma}(\Delta)$ is contingent upon how "error" is defined in a particular test application. If the imprecision of an individual's scores is measured over a hypothetical set of repeated testings with classically parallel test forms and the test is primarily used to rank order individuals, the appropriate error would be $\hat{\sigma}_E$. In this case the sampling of items does not introduce error into the estimation of an individual's domain score relative to the average domain score for the group (Brennan & Kane, 1977a, p.283). On the other hand, if the imprecision is measured over a hypothetical set of repeated testings with randomly parallel test forms and the test is used to determine an individual's level of performance in the domain irrespective of the performance of other students, either $S.E._{meas.}(\underline{x}_a)$ or $\hat{\sigma}(\Delta)$ would be appropriate. In that case the sampling of items does introduce error into the estimator of an individual's domain score. The final decision then rests with whether the individual specific or "blanket estimate" should be employed. Although the estimate of error for each individual P has several advantages which were mentioned previously, it is also usually subject to excessively large sampling fluctuations (cf. Lord, 1955). Therefore, when the standard error is applied to the cutting score, the individual specific estimate $S.E._{meas.}(\underline{x}_a)$

should be computed; otherwise, preference should be given to the estimate $\hat{\sigma}(\Delta)$ averaged over individuals.

Domain-referenced agreement index. The last group specific statistic Φ measures the contribution of the testing procedure to the dependability of the decisions that are made. In contrast to the other domain score estimation statistics, Φ can be viewed as an agreement index corrected for chance. Developed within the framework of generalizability theory, the Φ index has many of the features of its squared-error loss function counterpart $\Phi(\lambda)$. However, since it is not defined in relation to the cutting score (λ), it furnishes different information and answers a different question. While Φ is recommended as a general-purpose index of dependability for domain-referenced interpretations (Brennan, 1980, p.204), the specific utility of the index in practice needs further clarification.

Recommendations. All of the foregoing statistics except $\hat{\gamma}_i$ can be manually computed from just one test administration. This advantage plus their ease of interpretation and application suggest that they could be used by teachers to estimate the precision of an individual's scores on repeated testings with classically parallel or randomly parallel test forms. A teacher need only select the point estimate $\hat{P}$, one of the estimates of standard error ($S.E._{meas.}(\underline{x}_a)$, $\hat{\sigma}_E$, or $\hat{\sigma}(\Delta)$), or the agreement index Φ.

There are two problems, however, that appear to restrict the usefulness of those statistics for individual decision making at the classroom or school level. First, as noted in the preceding descriptions of the statistics, a relatively large number of items per objective or the total

number of items on the test must be used to obtain reliable estimates. This requirement is contrary to the characteristics of most teacher-made and professionally-developed criterion-referenced tests. Seldom are there more than 10 items per objective on these tests and the use of the total score for classroom level decisions about individuals is inconsistent with the intent and structure of the test. Total score decision making about an individual's level of competence in the domain is justified for individual certification decisions in the context of minimum competency testing programs. Those decisions, however, are rarely made independent of a competency standard. Second, the actual use of test scores for decision making typically requires some reference point or standard--relative or absolute. While the estimate of a student's domain score supplies information about how much a student knows or his/her level of skill in a given content domain, it is difficult to make placement, formative, or summative decisions based on that information alone. Once a score is transformed into a normative score (e.g., percentile, stanine) for norm-referenced test score interpretations or a score is referenced to a mastery standard for criterion-referenced test score interpretations, individual decision making becomes considerably easier. The aforementioned statistics must be assessed in light of these practices. Consequently, the meaning of $\hat{P}$ and the various errors of measurement might be enhanced if they are related to the cutting score as in the application of the Uncertainty Band concept. The Φ index has already received attention from this perspective in the form of the $\Phi(\lambda)$ squared-error loss agreement index.

These issues are not generalizable to estimates of the average domain score. The precision and utility of "averages" in the evaluation of competing instructional programs using appropriate units of analysis

(individual, class, school, or district) have been demonstrated (See, for example, Sirotnik, 1974). When criterion-referenced tests are employed in program evaluation, the use of various matrix sampling designs can provide more dependable estimates of average domain scores without any loss in the dependability of estimates of individual domain scores (cf. Brennan, 1980, p. 217).

Summary of Guidelines for "Reliability" Index Selection

The foregoing reviews of indices proposed as measures of criterion-referenced test "reliability" was organized into two sections: (1) a discussion of preliminary considerations related to the assumptions, interpretations, and uses of the indices and (2) an evaluation of the characteristics, precision, and practicability of the indices. The former was intended to provide a foundation for choosing the appropriate category of "reliability" for a particular application; the latter was designed to guide the selection of a specific index within the chosen category.

The first set of considerations and the information that is supplied germane to the choice of a "reliability" category are listed below:

1. Test forms assumption -- classically parallel or randomly parallel test forms;
2. Setting the cutting score -- cutting score and probabilities of false mastery and false nonmastery classification errors;
3. Test score interpretation and decision making -- score interpretation, type of decision, and seriousness of losses associated with decision errors.

The evaluation of the indices by category (See Tables 2, 3, and 4) suggested the following conclusions:

1. Threshold loss function indices -- Hambleton and Novick's (1973) method for estimating p_o is the easiest to understand, to compute, and to interpret; it appears to have the greatest utility for classroom test construction and decision making. Huynh's (1976) method for estimating p_o and κ is relatively precise and practicable; it should be considered by the more sophisticated test maker at the district or state level.
2. Squared-error loss function indices -- Livingston's (1972a) and Brennan's (1980) indices provide meaningful information about the consistency of scores in relation to the cutting score, particularly for placement tests. The choice between the indices depends upon the assumption about test forms.
3. Domain score estimation statistics -- The various estimates seem to have limited usefulness for classroom decisions unless they are linked to the cutting score, as exemplified by Millman's (1974b) approach. Estimates of the average domain score appear to have greater utility in the context of program evaluation decisions.

References

Algina, J., & Noe, M.J. A study of the accuracy of Subkoviak's single-administration estimate of the coefficient of agreement using two true-score estimates. Journal of Educational Measurement, 1978, 15, 101-110.

APA/AERA/NCME Committee. Standards for educational & psychological tests (Rev. Ed.). Washington, D. C.: American Psychological Association, 1974.

Antonak, R. F. A computer program to compute measures of response agreement for nominal scale data obtained from two judges. Behavior Research Methods & Instrumentation, 1977, 9, 553.

Berk, R. A. The relative merits of item transformations and the cloze procedure for the measurement of reading comprehension. Journal of Reading Behavior, 1979, 11, 129-138.

Berk, R. A. An extension of Huynh's decision-consistency reliability model to multipoint criterion-referenced test items. Paper presented at the annual meeting of the Eastern Educational Research Association, Norfolk, VA, March 1980.

Berk, R. A. A comparison of six content domain specification strategies for criterion-referenced tests. Educational Technology, in press. (a)

Berk, R. A. Estimation of test length for domain-referenced reading comprehension tests. Journal of Experimental Education, in press. (b)

Berk, R. A., & Campbell, K. L. A FORTRAN program for Cohen's kappa coefficient of observer agreement. Behavior Research Methods & Instrumentation, 1976, 8, 396.

Brennan, R. L. KR-21 and lower limits of an index of dependability for mastery tests (ACT Technical Bulletin No. 27). Iowa City, IA: American College Testing Program, December 1977.

Brennan, R. L. Handbook for GAPID: A FORTRAN IV computer program for generalizability analyses with single-facet designs (ACT Technical Bulletin No. 34). Iowa City, IA: American College Testing Program, October 1979.

Brennan, R. L. Applications of generalizability theory. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 186-232.

Brennan, R. L., & Kane, M. T. An index of dependability for mastery tests. Journal of Educational Measurement, 1977, 14, 277-289. (a)

Brennan, R. L., & Kane, M. T. Signal/noise ratios for domain-referenced tests. Psychometrika, 1977, 42, 609-625; Errata, 1978, 43, 289. (b)

Cardinet, J., Tourneur, Y., & Allal, L. The symmetry of generalizability theory: Applications to educational measurement. Journal of Educational Measurement, 1976, 13, 119-135.

Carver, R. P. Special problems in measuring change with psychometric devices. In Evaluative research: Strategies and methods. Pittsburgh, PA: American Institutes for Research, 1970. Pp. 48-63.

Cochran, W. G. Sampling techniques (2nd ed.). New York: Wiley, 1963.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 37-46.

Cronbach, L. J., Gleser, G. C., Nanda, H., & Rajaratnam, N. The dependability of behavioral measurements: Theory of generalizability for scores and profiles. New York: Wiley, 1972.

Erlich, O., & Shavelson, R. Generalizability of measures: A computer program for two- and three-facet designs. Behavior Research Methods & Instrumentation, 1976, 8, 407-408.

Fremer, J. Using and interpreting the results of criterion-referenced tests: Promise, progress, and unresolved problems. Invited paper presented at the first Johns Hopkins University National Symposium on Educational Research, Washington, D. C., October 1978.

Hambleton, R. K. Test score validity and standard-setting methods. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 80-123.

Hambleton, R. K., Hutton, L. R., & Swaminathan, H. A comparison of several methods for assessing student mastery in objectives-based instructional programs. Journal of Experimental Education, 1976, 45, 57-64.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Hambleton, R. K., Powell, S., & Eignor, D. R. Issues and methods for standard-setting. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1979.

Hambleton, R. K., Swaminathan, H., Algina, J., & Coulson, D. B. Criterion-referenced testing and measurement: A review of technical issues and developments. Review of Educational Research, 1978, 48, 1-47.

Harris, C. W. An interpretation of Livingston's reliability coefficient for criterion-referenced tests. Journal of Educational Measurement, 1972, 9, 27-29.

Harris, C. W., & Pearlman, A. P. Conventional significance tests and indices of agreement or association. In C. W. Harris, A. P. Pearlman, and R. R. Wilcox (Eds.), Achievement test items: Methods of study (CSE Monograph Series in Evaluation, No. 6). Los Angeles: Center for the Study of Evaluation, University of California, 1977. Pp. 17-44.

Huynh, H. On the reliability of decisions in domain-referenced testing. Journal of Educational Measurement, 1976, 13, 253-264.

Huynh, H. The kappamax reliability index for decisions in domain-referenced testing. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Huynh, H. Statistical inference for two reliability indices in mastery testing based on the beta-binomial model. Journal of Educational Statistics, 1979, 4, 231-246.

Huynh, H., & Saunders, J. C. Accuracy of two procedures for estimating reliability of mastery tests. Paper presented at the annual meeting of the Eastern Educational Research Association, Kiawah Island, SC, February 1979.

Jackson, P. H. Simple approximations in the estimation of many parameters. British Journal of Mathematical and Statistical Psychology, 1972, 25, 213-229.

Kane, M. T., & Brennan, R. L. Agreement coefficients as indices of dependability for domain-referenced tests. Applied Psychological Measurement, in press.

Keats, J. A., & Lord, F. M. A theoretical distribution for mental test scores. Psychometrika, 1962, 27, 59-72.

Lewis, C., Wang, M. M., & Novick, M. R. Marginal distributions for the estimation of proportions in m groups. Psychometrika, 1975, 40, 63-75.

Linn, R. L. Issues of reliability in measurement for competency-based programs. In M. A. Bunda and J. R. Sanders (Eds.), Practices and problems in competency-based measurement. Washington, D. C.: National Council on Measurement in Education, 1979. Pp. 90-107.

Livingston, S. A. Criterion-referenced applications of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26. (a)

Livingston, S. A. A reply to Harris' "An interpretation of Livingston's reliability coefficient for criterion-referenced tests." Journal of Educational Measurement, 1972, 9, 31. (b)

Livingston, S. A. Reply to Shavelson, Block, and Ravitch's "Criterion referenced testing: Comments on reliability." Journal of Educational Measurement, 1972, 9, 139-140. (c)

Livingston, S. A. Psychometric techniques for criterion-referenced testing and behavioral assessment. In J. D. Cone and R. P. Hawkins (Eds.), Behavioral assessment: New directions in clinical psychology. New York: Brunner/Mazel, 1977. Pp. 308-329.

Livingston, S. A., & Wingersky, M. S. Assessing the reliability of tests used to make pass/fail decisions. Journal of Educational Measurement, 1979, 16, 247-260.

Lord, F. M. Sampling fluctuations resulting from the sampling of test items. Psychometrika, 1955, 22, 1-22.

Lord, F. M. Do tests of the same length have the same standard error of measurement? Educational and Psychological Measurement, 1957, 17, 510-521.

Lord, F. M. Tests of the same length do have the same standard error of measurement. Educational and Psychological Measurement, 1959, 19, 233-239.

Lord, F. M., & Novick, M. R. Statistical theories of mental test scores. Reading, MA: Addison-Wesley, 1968.

Marascuilo, L. A. Statistical methods for behavioral science research. New York: McGraw-Hill, 1971.

Marshall, J. L. The mean split-half coefficient of agreement and its relation to other test indices: A study based on simulated data (Technical Report No. 350). Madison, WI: Wisconsin Research and Development Center for Cognitive Learning, 1976.

Marshall, J. L., & Serlin, R. C. Characteristics of four mastery test reliability indices: Influence of distribution shape and cutting score. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, April 1979.

Millman, J. Criterion-referenced measurement. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan, 1974. Pp. 311-397. (a)

Millman, J. Sampling plans for domain-referenced tests. Educational Technology, 1974, 14, 17-21. (b)

Millman, J. Computer-based item generation. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 32-43.

Novick, M. R., Lewis, C., & Jackson, P. H. The estimation of proportions in m groups. Psychometrika, 1973, 38, 19-45.

Popham, W. J. Domain specification strategies. In R. A. Berk (Ed.) Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 15-31.

Reid, J. B., & Roberts, D. M. A Monte Carlo comparison of phi and kappa as measures of criterion-referenced reliability. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Safrit, M. J., & Stamm, C. L. The reliability of criterion-referenced measures of motor behavior: A comparative study. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, April 1979.

Shavelson, R. J., Block, J. H., & Ravitch, M. M. Criterion-referenced testing: Comments on reliability. Journal of Educational Measurement, 1972, 9, 133-137.

Shepard, L. A. Technical issues in minimum competency testing. In D. C. Berliner (Ed.), Review of research in education (Vol. 8). Itasca, IL: F. E. Peacock, 1980.

Shoemaker, D. M. Principles and procedures of multiple matrix sampling. Cambridge, MA: Ballinger, 1973.

Sirotnik, K. A. Introduction to matrix sampling for the practitioner. In W. J. Popham (Ed.), Evaluation in education: Current applications. Berkeley, CA: McCutchan, 1974. Pp. 451-529.

Subkoviak, M. J. Estimating reliability from a single administration of a mastery test. Journal of Educational Measurement, 1976, 13, 265-276.

Subkoviak, M. J. Empirical investigation of procedures for estimating reliability for mastery tests. Journal of Educational Measurement, 1978, 15, 111-116.

Subkoviak, M. J. Decision-consistency approaches. In R. A. Berk (Ed.), Criterion-referenced measurement: The state of the art. Baltimore, MD: The Johns Hopkins University Press, 1980. Pp. 129-185.

Swaminathan, J., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: A decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

Swineford, F. Note on "Tests of the same length do have the same standard error of measurement." Educational and Psychological Measurement, 1959, 19, 241-242.

Thorndike, R. L. Reliability. In A. Anastasi (Ed.), Testing problems in perspective. Washington, D. C.: American Council on Education, 1967. Pp. 284-291.

Wixon, D. R. Cohen's kappa coefficient of observer agreement: A BASIC program for minicomputers. Behavior Research Methods & Instrumentation, 1979, 11, 602.

Table 1

Score and Decision Characteristics for Three Categories of "Reliability"

Characteristic	Category of "Reliability"		
	Threshold Loss	Squared-Error Loss	Domain Score Estimation
Score interpretation	Individual score is referenced to cutting score	Individual score is referenced to cutting score	Individual score is used to estimate domain score
Type of decision or information required for decision	Mastery or nonmastery classification	Degree of mastery or nonmastery along score continuum	Level of competency in content domain
Losses associated with decision errors	Losses are equally serious	Losses related to misclassified students who are far above or below cutting score are more serious than losses related to other misclassified students	Consideration of losses varies with the estimation procedures

Table 2
Evaluation of Threshold Loss Function Indices
(Listed in order of increasing overall complexity)

Index	Definition	Source	No. of Administrations	Test Forms Assumption	Distribution Assumptions	Correction for Chance Agreement	Comments
2 Percentages	Percentage of masters on one test form compared to percentage of masters on a parallel form; Percentage of masters in one group on a single form compared to percentage of masters in a comparable group on a second administration of the same form	Carver (1970)	2	1 Form or Parallel	None	No; No	Advantage: Manually calculable. Disadvantages: Comparison of two percentages based on one test form or two parallel forms is insensitive to the consistency of individual classification decisions (Subkoviak, 1980); it is not a measure of agreement.
$\hat{p}_o$	Proportion of individuals consistently classified as masters and nonmasters on repeated measurements of one form or parallel forms	Hambleton and Novick (1973)	2	1 Form or Parallel	None	No	Advantages: Provides unbiased estimate of p_o. Manually calculable and easily interpretable. Disadvantages: Requires two test administrations. When only one test form is used, testing effect can spuriously deflate estimates. $\hat{p}_o$ has comparatively large standard errors for classroom size samples (Subkoviak, 1978, 1980).
$\hat{\kappa}$	Proportion of individuals consistently classified as above beyond that expected by chance	Swaminathan, Hambleton, and Algina (1974) based on Cohen (1960)	2	1 Form or Parallel	None	Yes[a]	Advantages: Manually calculable; computer programs are also available (e.g., Antonak, 1977; Berk & Campbell, 1976; Mixon, 1979). κ and ϕ coefficients are virtually identical (Marshall & Berlin, 1979; Reid & Roberts, 1978). Disadvantages: Yields biased estimate of κ. Sensitivity to test length and variance may restrict range of values. Requires two test administrations. Testing effect based on test-retest design can spuriously deflate estimates. Dependence of index values on cutting score, marginal frequencies, test length, and score variance complicates index interpretation.

continued

[a]The κ index is corrected for chance according to the marginal frequencies of the particular contingency table data set under consideration. This constraint may be viewed as a disadvantage of using κ for tests intended for individual decisions at the classroom level.

Table 3 (continued)

Index	Definition	Source	No. of Administrations	Test Forms Assumption	Distribution Assumptions	Correction for Chance Agreement	Comments
$\hat{S}(\hat{p}_o)$	Same as $\hat{p}_o$ defined above	Marshall (1976)	1	Parallel	Binomial distribution of observed scores over repeated measurements (assumes 0-1 items, statistical independence of items, and equal item difficulties)	No	Average p_o across all possible split-halves of a single test, extended to a full-length (double-length test) estimate with Spearman-Brown type prophesy formula. Advantages: Requires only one test administration. Comparatively small standard errors for classroom-size samples (Subkoviak, 1978, 1980). Easily interpretable. Disadvantages: Yields biased estimates for short tests (Subkoviak, 1978, 1980). Provides no estimate of κ. Computationally complex with no readily available computer program.
$\hat{p}_o$, $\hat{\kappa}$	Same as $\hat{p}_o$ and $\hat{\kappa}$ defined above	Subkoviak (1976, 1980)	1	Parallel	Binomial or compound binomial distribution of observed scores over repeated measurements	No; Yes[a]	Observed proportion correct scores and KR-20 coefficient are used to obtain linear regression approximations to the compound binomial p_o. Relies on the estimation of the true score for each person. Consistency index for each person is averaged over the population to produce $\hat{p}_o$. Advantages: Same as above. Disadvantages: Yields biased estimates of p_o for short tests (Algina & Noe, 1978; Subkoviak, 1978). Sensitivity of $\hat{\kappa}$ to test length and variance may restrict range of values. Computationally complex with no readily available computer program.

continued

[a] The κ index is corrected for chance according to the marginal frequencies of the particular contingency table data set under consideration. This constraint may be viewed as a disadvantage of using κ for tests intended for individual decisions at the classroom level.

Table 2 (continued)

Index	Definition	Source	No. of Administrations	Test Forms Assumption	Distribution Assumptions	Correction for Chance Agreement	Comments
$\hat{p}_o$; $\hat{\kappa}$	Same as p_o and κ defined above	Huynh (1976)	1	Parallel[b]	Binomial distribution of observed scores over repeated measurements Beta distribution of true scores	No; Yes[a]	Indices computed from the univariate and bivariate distributions. Under similar distribution assumptions for double-length tests p_o equals $\hat{s}(p_o)$ (Marshall & Serlin, 1979). Advantages: Same as above. Based on the mathematically elegant Keats and Lord (1962) model. Provides most accurate estimates of p_o for unimodal distributions. Violation of equal item difficulty assumption seems to have negligible effect on estimates (Huynh & Saunders, 1979; Subkoviak, 1978). Tabled values of p_o and κ for tests with five to 10 items per objective as well as a computer program are currently available (Huynh, 1979; Subkoviak, 1980). Disadvantages: Yields biased (slightly conservative) estimates of p_o and κ for short tests (Huynh & Saunders, 1979; Subkoviak, 1978). Sensitivity of κ to test length and variance may restrict range of values. Most conceptually and mathematically complex approach.
KH	Kappamax (maximum value of κ)	Huynh (1977)	1	Parallel[b]	Same as above	Yes[a]	Upperbound of κ is usually reached at a cutting score near the mean, although it is emphasized that KH is not a function of the cutting score. It reflects the situation in which a test functions best in terms of decision consistency. Advantages: Same as above. Disadvantages: Same as above. Interpretation and utility of KH by itself or in conjunction with κ are unclear.

[a] The κ index is corrected for chance according to the marginal frequencies of the particular contingency table data set under consideration. This constraint may be viewed as a disadvantage of using κ for tests intended for individual decisions at the classroom level.

[b] Huynh (1976, pp. 254-55, 1977, p.1) defined two equivalent test forms as two independent random samples of items drawn from a specified item universe. While these test forms could be randomly parallel, the author indicates that they possess the properties of classically parallel forms essential for the assumptions of the beta-binomial model (Huynh, personal communication, November 1979).

Table 3

Evaluation of Squared-Error Loss Function Indices
(Both indices are rated equal in overall complexity)

Index	Definition	Source	No. of Administrations	Test Forms Assumption	Distribution Assumptions	Correction for Chance Agreement	Comments
$k^2(X,\mathcal{T}_x)$	Ratio of true score to expected observed score mean squared deviations from the cutting score--error variance $\sigma^2(\delta)$ based on person X item interaction	Livingston (1972a, 1977)	1	Parallel	None	No	A generalization of the classical reliability coefficient, where the expected mean squared deviation about the cutting score is substituted for the variance about the mean score. When the cutting score equals the mean, k^2 reduces to the KR-20 coefficient. Advantages: Requires only one test administration. Manually calculable. Disadvantage: Dependence of index values on cutting score, test length, and the estimated true or universe score variance and error variance complicates index interpretation.
$\Phi(\lambda)$	Ratio of universe score to expected observed score mean squared deviations from the cutting score--error variance $\sigma^2(\Delta)$ based on item effect plus person X item interaction	Brennan (1980); Brennan and Kane (1977a, 1977b)	1	Randomly parallel	None	No	Measures the dependability of mastery-nonmastery decisions based on the testing procedure[a]. Derived from generalizability theory (Cronbach et al., 1972). When the cutting score (λ) equals the mean, $\Phi(\lambda)$ is identical to the KR-21 coefficient (Brennan, 1977). Advantages: Same as above. Computer programs especially useful for more complex analysis of various designs and the estimation of variance components are available (e.g., Brennan, 1979; Erlich & Shavelson, 1976). Estimation procedures are very flexible to accommodate different test structures and score uses. Disadvantage: Same as above.

[a]Brennan (1980, pp. 204-213) also defined another index of dependability Φ that is independent of any cutting score. It estimates the contribution of the testing procedure to the dependability of the decisions, over what would be expected on the basis of chance agreement. Unlike the above indices, the Φ index is most useful as a domain score estimation statistic (See Table 4).

Table 4

Evaluation of Domain Score Estimation Statistics
(Statistics within each category are listed in order of increasing overall complexity)

	Statistic	Definition	Source	No. of Administrations	Test Forms Assumption	Comments
Individual Specific	$\hat{P}$	Point estimate of proportion of items answered correctly in item domain	Hambleton et al. (1978)	1	None	Advantages: Provides unbiased estimate of P. Manually calculable and easily interpretable. Disadvantage: Estimate is highly unreliable for small numbers of items related to objectives (Hambleton, Hutton, & Swaminathan, 1976).
	$\hat{\epsilon}_p$	Standard error of a proportion--binomial formula for sampling error	Berk (in press a); Cochran (1963, pp. 49-59); Millman (1974a, 1974b)	1	Randomly parallel	Advantages: Provides unbiased estimate of error based on infinite or finite item domain; latter might require finite population correction (fpc) if item sample is large relative to item domain. Manually calculable and easily interpretable; tabled values are also widely available (e.g., Marascuilo, 1971). Standard error remains the same for all tests of the same length irrespective of item format (Lord, 1957). Error can be used to set up a confidence interval or Uncertainty Band (UB) around an individual's $\hat{P}$ or the P standard for mastery. Disadvantages: Same as above. Computed using only $\hat{P}$ and the number of items; estimate does not depend on the content, form, or statistical characteristics of the items. Therefore, estimates of error for a given test length will be lower for extreme values of P and higher for moderate values. Manual computation is tedious because the error must be computed for each individual $\hat{P}$.
	S.E.$_{\text{meas.}}(x_p)$	Standard error of measurement--binomial formula for sampling error	Lord (1955, 1957, 1959)	1	Randomly parallel	Equivalent to the unbiased estimate of ϵ_p based on an infinite item domain (See Cochran, 1963, p. 51). Advantages: Same as above. Disadvantages: Same as above.
	$\hat{Y}_i$	Modified marginal mean (quasi-Bayesian) point estimate of P	Hambleton, Hutton, and Swaminathan (1976)	1	None	Extends the binomial model underlying $\hat{\epsilon}_p$ and S.E.$_{\text{meas.}}(x_p)$ by incorporating direct ($\hat{P}$), collateral (group performance), and prior information (past test performance) into the estimate of P. Advantages: Produces smallest absolute error compared to P, classical model II estimate (Jackson, 1972), Bayesian joint modal estimate (Novick, Lewis, & Jackson, 1973), and marginal mean esti-

continued

Table 4 (continued)

	Statistic	Definition	Source	No. of Administrations	Test Forms Assumption	Comments
						mate (Lewis, Wang & Novick, 1975). Estimation procedures seem to be effective for subtests with as few as 8 items and for homogeneous score distributions. Disadvantages: Conceptually and computationally complex. Exact specification of mean and variance of prior distribution is difficult. Distribution assumptions may not be upheld in practice.
Group Specific	$\hat{\sigma}_E$	Standard error of measurement--average of individual specific errors of measurement $\hat{\sigma}(E_{pa})$	Lord and Novick (1968, pp. 60, 153-160) (Also see Brennan, 1980)	1	Parallel	Conventional estimate of error derived from classical reliability theory. Equivalent to $\hat{\sigma}(\Delta)$ in the context of generalizability theory, which is based on person X item interaction. Advantages: Manually calculable and easily interpretable. Error can be used to set up a confidence interval around an individual's P. Disadvantages: Yields biased estimate of error (See Lord & Novick, 1968, p. 152). Estimate is unreliable for short subtests. Due to the fluctuation of the individual specific standard errors with different values of $\hat{P}$, the average of those errors will tend to be lowest when there is a large number of extremely high and/or extremely low scores in the distribution (e.g., skewed or bimodal distribution); that low estimate can be very misleading.
	$\hat{\sigma}(\Delta)$	Standard error of measurement--average of individual specific errors of measurement S.E.$_{\text{meas.}}(\underline{x}_p)$	Brennan (1980)	1	Randomly parallel	Incorporates the effect due to item sampling along with the person X item interaction into the estimate. Derived from generalizability theory (Cronbach et al., 1972). Advantages: Same as above. For the simple matrix sampling of persons and items, $\hat{\sigma}(\Delta)$ is equivalent to the square root of Lord and Novick's (1968, p.251) unbiased estimate $\hat{\sigma}^2_{\bar{x}}(n_i)$; therefore, it can be computed easily from the mean and variance using formula (11.9.4). Disadvantages: Same as above, except for first statement.
	Φ	Ratio of universe score variance to expected observed score variance--error variance $\sigma^2(\Delta)$ based on item effect plus person X item interaction	Brennan (1980)	1	Randomly parallel	Measures the contribution of the testing procedure to the dependability of the decisions. Derived from generalizability theory (Cronbach et al., 1972). Φ has an upper limit equal to the KR-20 coefficient and a lower limit equal to the KR-21 coefficient (Brennan, 1980, p. 213). Advantages: Manually calculable; computer programs are also available (e.g., Brennan, 1979). Estimation procedures are very flexible to accommodate different test structures and score uses. Disadvantages: Dependence of index values on test length and the estimated universe score and error variance complicates index interpretation. Specific application of the index in practice is unclear.